CHIST-ERA HLU Master School: ATLANTIS

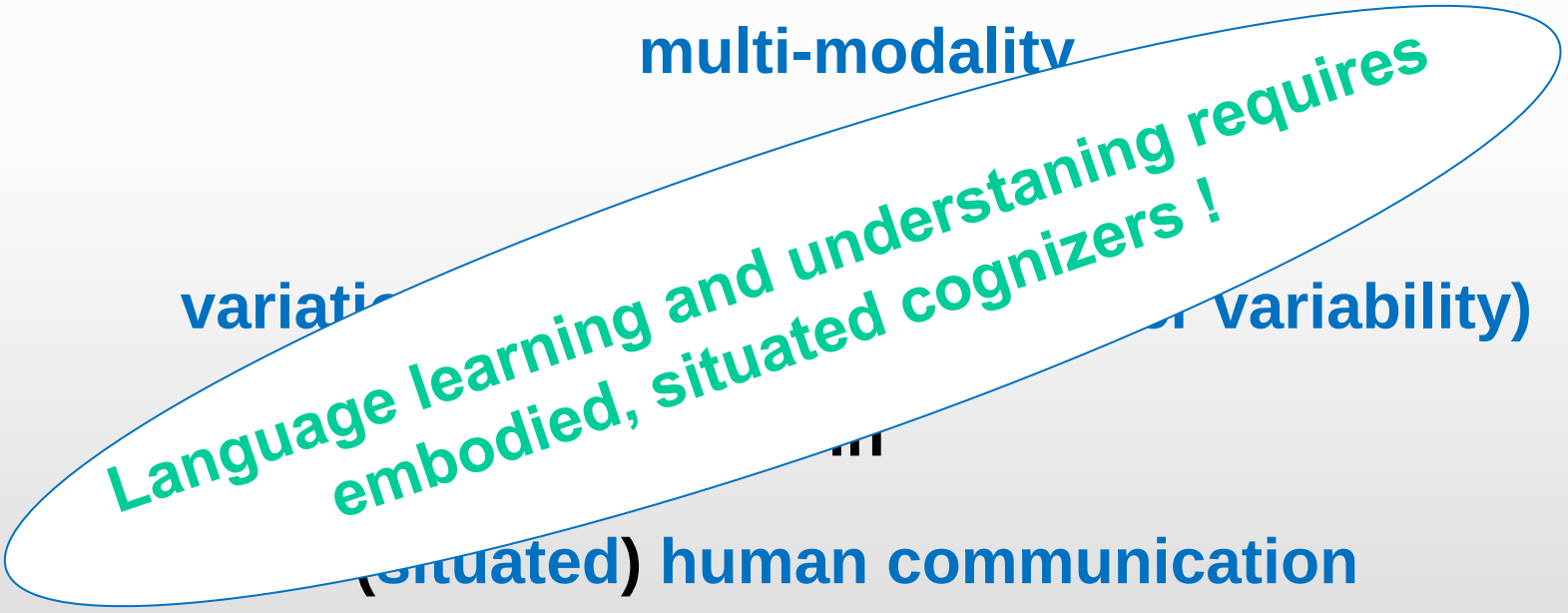
Multimodal Corpora for Crossmodal, Grounded Language Learning

Brigitte Krenn, Friedrich Neubarth

OFAI, Vienna, Austria

Overall Goal

To better understand



Robots

For natural HRI, robots have to be able to

- share representations of concepts with their communication partner

  (e.g. Kruijff et al. 2010, Coradeschi et al. 2013)

- identify human communicative cues and extract and merge information transmitted via different channels

  (e.g. Scheutz et al. 2013, Kopp et al. 2013, Hüwel et al. 2006, Lemaignan et al. 2012)

Data collection

- Overall goal
  - develop and implement mechanisms to account for the multi-modal complexity of human communication
- Specific focus
  - multi-modality in task descriptions
- Means
  - data collection, analysis, interpretation and transformation of insights into learning games

Steps

- Research questions
- Design of data collection experiments
- Technical setup & recordings
- From data collection to annotated corpus

General research questions

- **Q1** – Which phenomena occur during task descriptions and what is their impact on comprehension?
- **Q2** – What is the inter- and intra-speaker variability in conveying respective information? Interaction Studies
- **Q3** – On which channels is relevant information transmitted? Interaction Studies
- **Q4** – What are the differences in how a task is transmitted between HH and HR dyads?

Q1,4 discussed in Schreitter S., Krenn B.: Exploring Inter- and Intra-speaker Variability in Multi-modal Task Descriptions, Proceedings of the 17th IEEE International Symposium on the Robot and Human Interactive Communication, (Ro-Man 2014), 2014

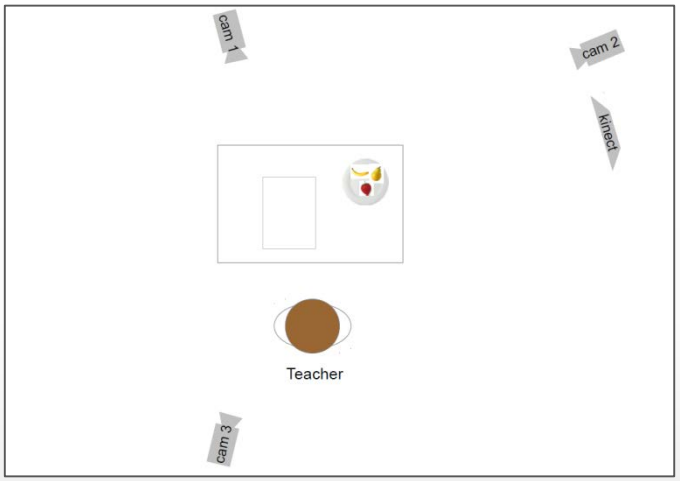
Q2,3 discussed in Gross S., Krenn B., Scheutz M.: Multi-modal referring expressions in human-human task descriptions and their implications for human-robot interaction, Interaction Studies (accepted on 19 Jan 2016).

The OFAI Multi-Modal Task Description Corpus
Data collection and analysis

Gross, S., Krenn, B.: The OFAI Multi-Modal Task Description Corpus. LREC 2016

Task 1: Arranging Fruits

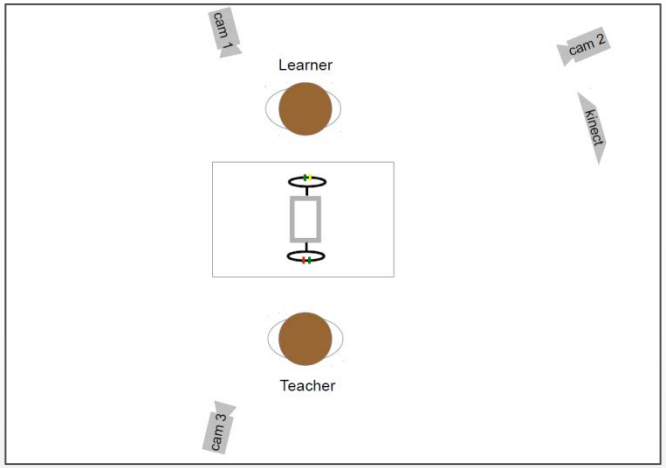
- **Scenario**
  - wooden fruits are arranged and re-arranged on a table
- **Focus**
  - investigating auditory cues of information structure (prosody, givenness, focus of attention)
  - voiced object names (*banana*, *strawberry, pear*)
- **Instructor**
  - performs and explains task



- Resulting dataset
  - multi-modal data from 22 humans

Task 2: collaboratively moving an object

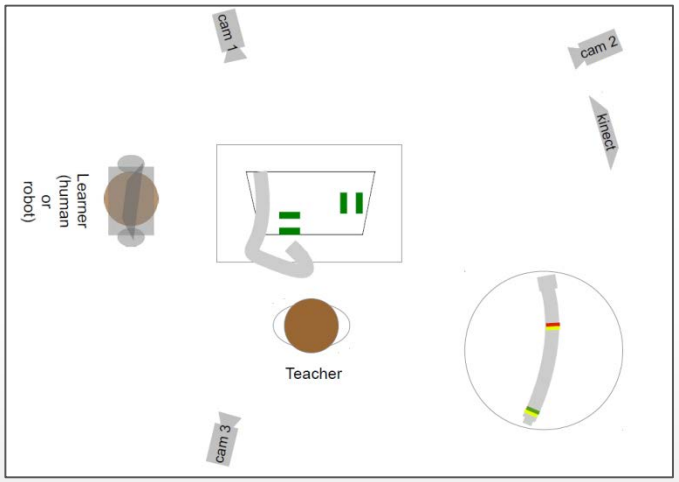
- **Scenario**
  - instructor and learner collaboratively move a board
- **Focus**
  - collaborative handling of single object
- **Instructor-learner pairs**
  - collaboratively move object
- **Instructor**
  - explains what to do
  - observes and when necessary corrects learner actions



- Resulting dataset
  - multi-modal interactions from 22 human-human pairs

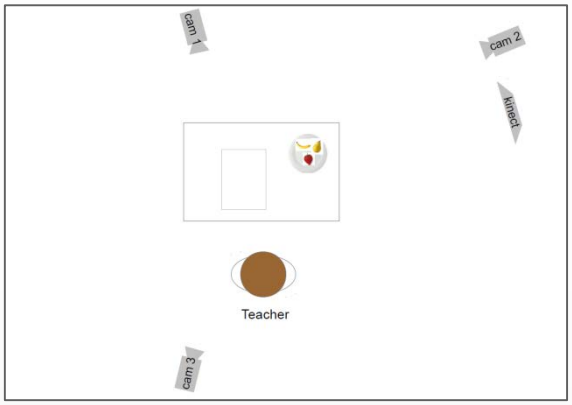
Task 3: mounting a tube

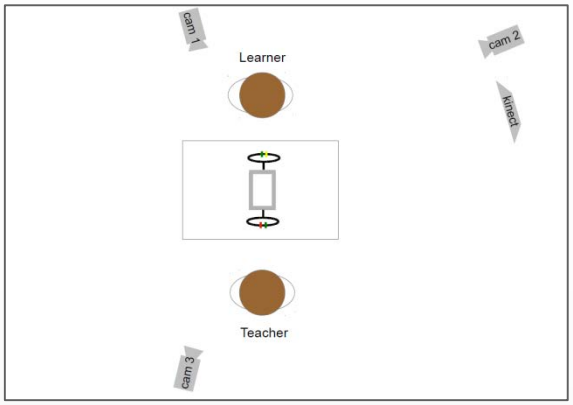
- **Scenario**
  - teacher explains and shows to a learner how to connect two separate parts of a tube and then to mount the tube in a box with holdings
- **Focus**
  - variation in object references
  - cues to direct the learner's attention
- **Instructor**
  - performs and explains the task

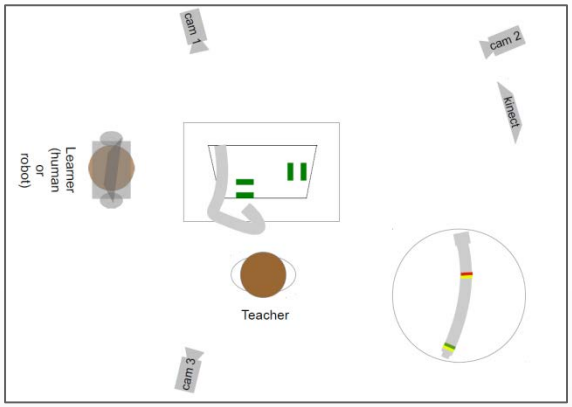


- Resulting datasets
  - multi-modal data from from 16 human-human pairs
  - 6 human-robot pairs

Task description experiments







- audio recordings of (teacher) utterances
  - wireless microphone worn by the instructor,
  - a receiver,
  - a sound mixer connected to a laptop,
  - Audacity for recording http://audacity.sourceforge.net/
- 3 videos:
  - teacher frontal, learner frontal, overall setting
- motion data
  - Qualisys System http://www.qualisys.com/
  - Kinect
- force data

Resulting data sets per task

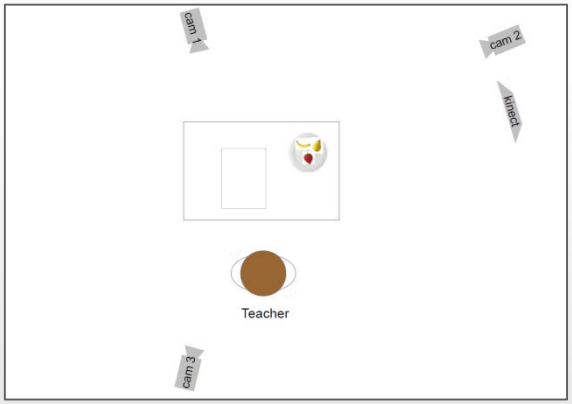
From:

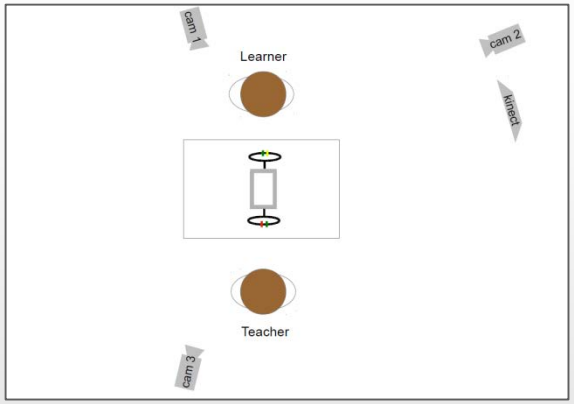
22 humans

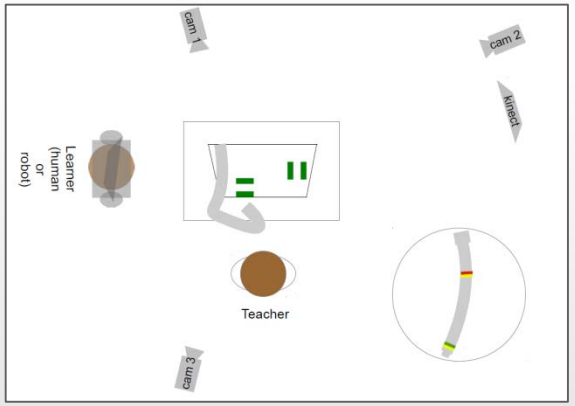
22 human-human pairs

16 human-human

6 human-robot pairs







Instructors/learners: 15 male, 7 female; av. age 27 years

Data Issues:

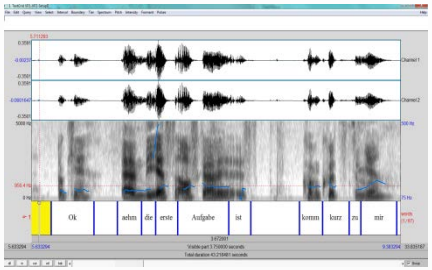
- **Size of data set:**
  - HHI: only 16 or 22 instructor-learner dyads per task
  - each task serves different purposes, therefore data from different tasks cannot used together for general purposes
  - sufficient for qualitative analysis, but too small to employ statistical tests
- **Generalisation between tasks:**
  - number and thus saliency of objects (e.g. Board vs. Tube and Holdings)
  - who acts, who observes (collaborative vs. teacher explains learner listens)
- **Familiarity:**
  - not all of the instructor-learner dyads knew each other before
- **Participants:**
  - are not balanced wrt. gender, age and education
  - Students or people working at the university
  - More male than female participants (HHI: 15:7, 12:4)

Sample Videos and Annotations

Data annotation: Tools

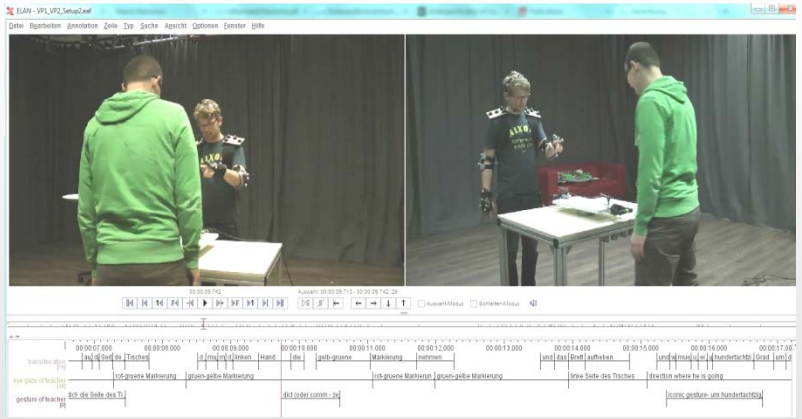
- Praat for
  - transcription of the audio files

  http://www.fon.hum.uva.nl/praat/



- ELAN for
  - synchronisation purposes and
  - Annotation
  - Tiers as csv → further processing

  https://tla.mpi.nl/tools/tla-tools/elan/

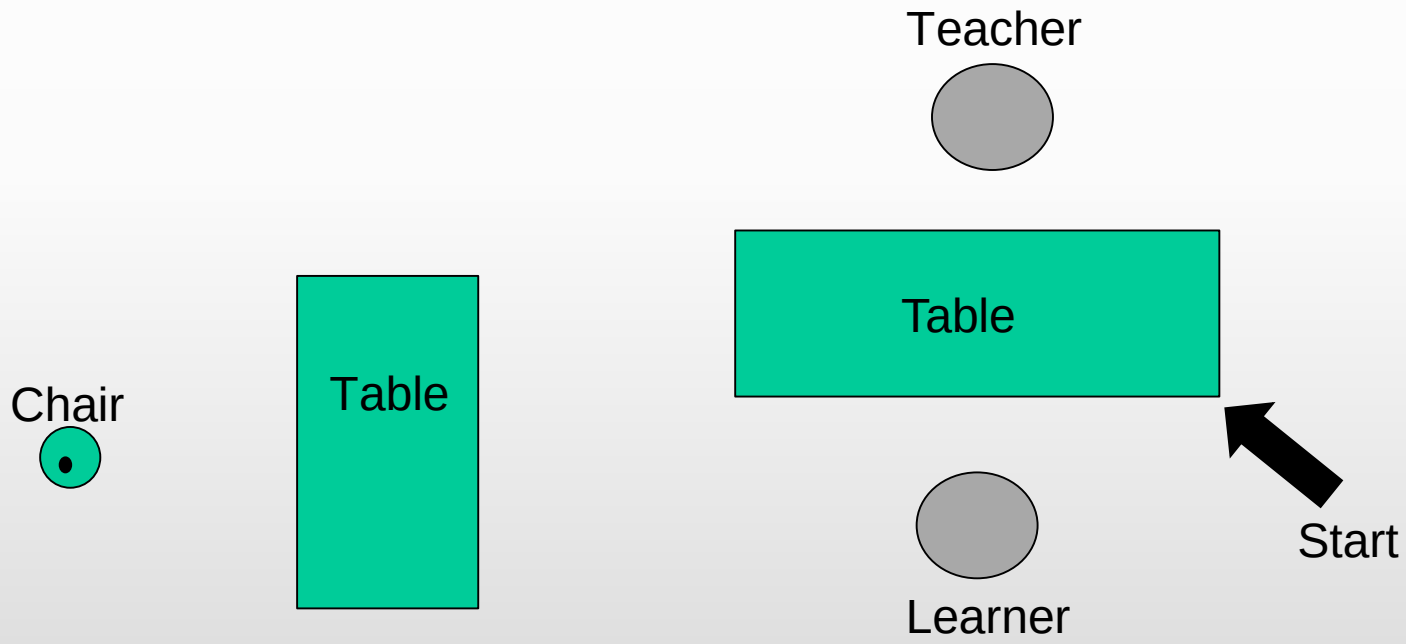


- TreeTagger for PoS-tags

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In Proceedings of the ACL SIGDAT-Workshop.

Let's do …

- New recordings for a navigation task
- Basic scenario



Technical preparation

- convert MTS file to AVI
  - any video converter

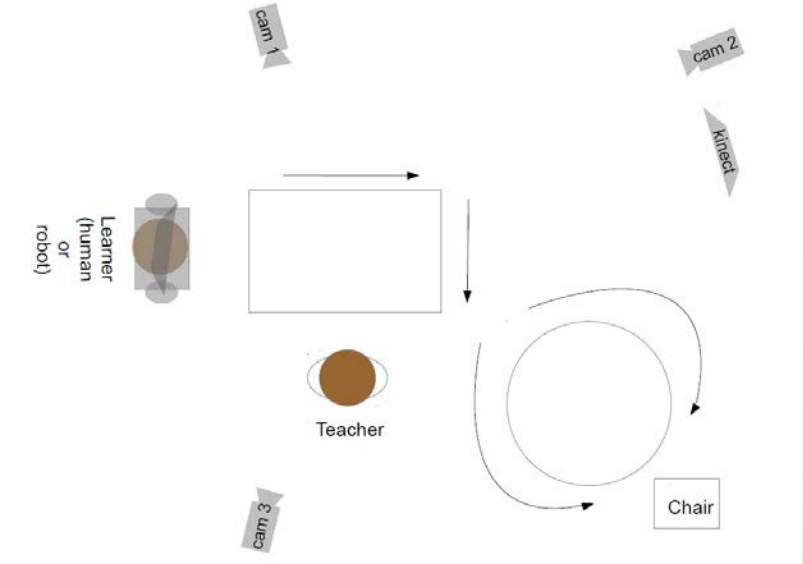
    http://www.any-video-converter.com/it/any-video-converter-free.php
- (trim videos, e.g. use any)
- load video into elan, create annotation file (.eaf), create annotation tiers, do annotations
  - https://tla.mpi.nl/tools/tla-tools/elan/

    http://fave.ling.upenn.edu/downloads/ELAN_Introduction.pdf

OFAI MMTD Corpus Task 4: spatial navigation

Spatial Navigation Task

- **Scenario**
  - teacher instructs learner which path to go
- **Focus**
  - Function of multi-modal cues in navigation instruction
  - corrections and redirections
- **Instructor**
  - explains task
- **Learner**
  - performs the task



- **Resulting dataset**
  - multi-modal data from from 16 human-human pairs
  - 6 human-robot pairs

Some observations*

- Gesture and speech
  - gesture is typically accompanied by speech
- Gesture and eye gaze – eye gaze is:
  - used by the teacher to be aware of whether the learner is following the instructions
  - lined up with gesture, e.g. for all gestures relating to a corner of the table, the eye gaze was directed at the following objects:

'Lerner', 12	'andere obere Ecke vom Tisch', 2
'obere Ecke vom Tisch', 10	'wandernde Blickrichtung', 1
'Ecke vom Tisch', 8	'kurze Tischkante', 1
'runder Tisch', 5	'lange Tischkante', 1

*This work was mainly carried out by Benjamin Fischer at OFAI.

Annotation Tiers

- Transcription of utterances
- Transliteration (speech normalized)
- PoS Tags
- Eye gaze of the teacher
- Gesture of the teacher
- Relevant objects
- Phrase boundary
- Prominence level
- Prosody

Perspective Taking:

the use of personal pronouns (*ich*, *du*, *wir*)

in the MMTD Corpus

Perspective Taking

Task Characteristics

Task 1 – object manipulation; active role of instructor; to video camera

Task 2 – object manipulation; collaborative task; active role of instructor & learner

Task 3 – object manipulation; active role of instructor; passive role of learner

Task 4 – navigation; passive role of instructor; active role of learner

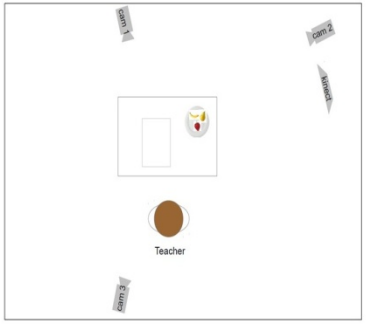
Investigate

prounouns as means of perspective taking

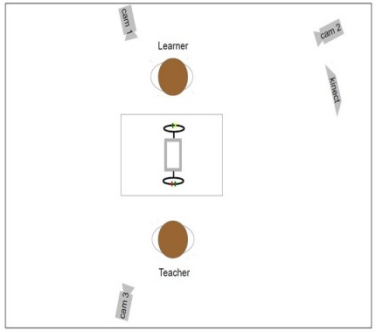
Investigate the use of *ich, du, wir* in task-oriented discourse

Literal versus impersonal *du/wir*

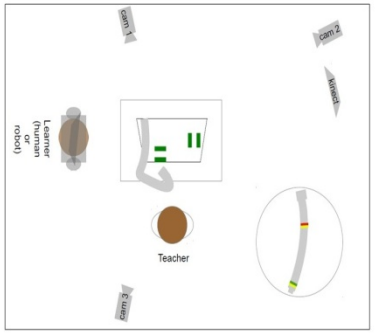
HH versus HR-dyads



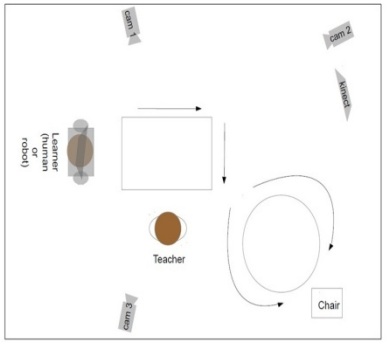
Task 1: arranging fruit
(datasets from 22 humans)



Task 2: collaboratively moving an object
(datasets from 22 human-human pairs)



Task 3: mounting a tube
(datasets from 16 human-human and 6 human-robot pairs)



Task 4: navigation
(datasets from 16 human-human and 6 human-robot pairs)

Perspective Taking: Results

- **Task Dependent Usage**

Active learner involvement

→ literal use of *ich, du, wir*

Only instructor conducts task

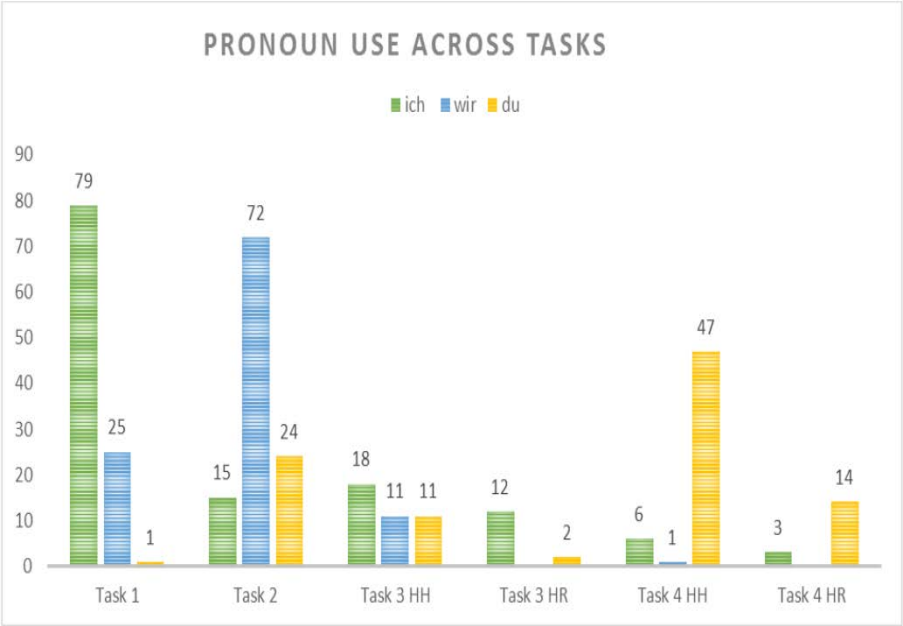
→ mixed use (literal vs impersonal)

→ visual and verbal cues necessary for disambiguation, e.g., literal *du* + pointing + *jetzt* , *bitte*

- **Differences in HH and HR Communication**

in HR: almost no *wir*

in HH & HR: Parallels in the use of *ich, du*



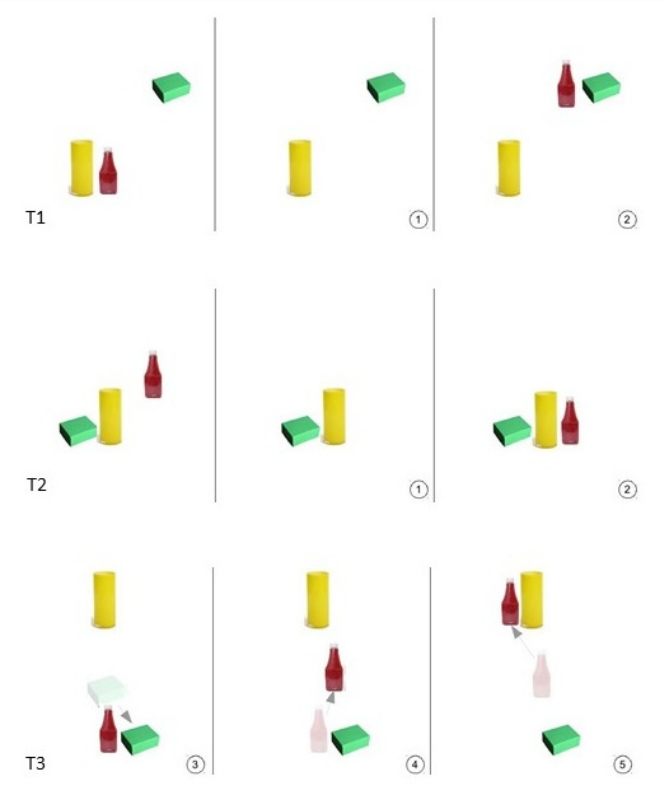
Krenn B., Gross S., Nussbaumer L.: Who Has to Do It? The Use of Personal Pronouns in Human-Human and Human-Robot-Interaction, *1st International Workshop on Investigating Social Interactions with Artificial Agents co-located with ICMI 2017. Glasgow, Scotland.*, 2017.

Overall: in HR-dyads humans were more explicit about who is supposed to do what

Action Verb Corpus

- Simple basic actions involving a small set of objects with verbal descriptions
- 3 Actions: TAKE + PUT, MOVE
- 3 Objects: BOX, BOTTLE, CAN
- Video, Kinect and LibMotion data
- Audio recodring (wav)
- Goal: incremental crossmodal (grounded) word learning

Action Verb Corpus – Summer Experiments



Task	Number of Recordings	Number of Actions per Recording
T1	20	4 TAKE/PUT-actions
T2	15	4 TAKE/PUT-actions
T3	11	10 PUSH-actions

AVC actions only	AVC actions and related objects
nehme – TAKE schiebe – PUSH stelle – PUT	nehme – TAKE schiebe – PUSH stelle – PUT dose – PRINGLES flasche – KETCHUP schachtel – TEAHORIZONTAL

— LREC 2018

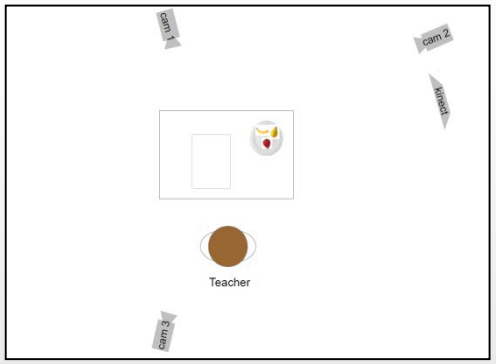
Stephanie Gross, Matthias Hirschmanner, Brigitte Krenn, Friedrich Neubarth, Michael Zillich. Action Verb Corpus

Crossmodal word learning

Crossmodal word learning

MMTD Corpus

Task 1: arranging fruit (datasets from 22 humans)



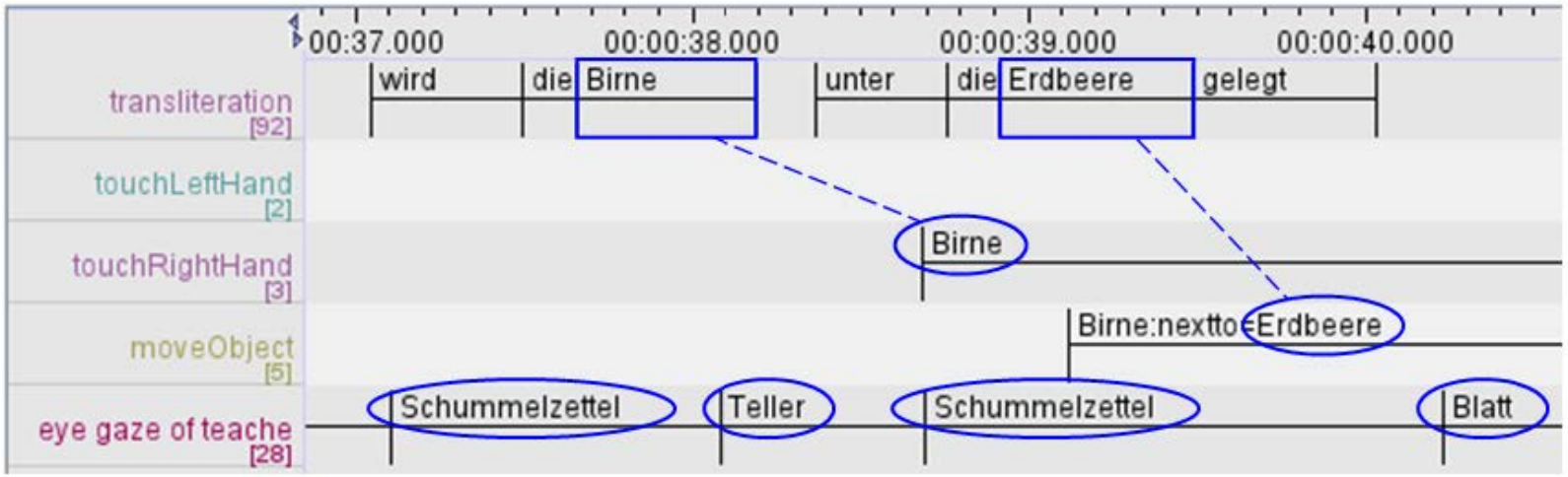
Visual Cues

VC1 – BIRNE
VC2 – BIRNE, ERDBEERE
VC3 – SCHUMMELZETTEL, TELLER, BLATT
VC4 – BIRNE, ERDBEERE, SCHUMMELZETTEL, TELLER, BLATT

Krenn B., Trapp M., Gross S., Neubarth F.: Crossmodal Cross-situational Learning with Attention, *IEEE ICDL-EPIROB 2017, Workshop on Computational Models for Crossmodal Learning. Lisbon, Portugal.*, 2017.

Crossmodal word learning

- **Crossmodal Relations**



Crossmodal word learning

- **Models and Results (Batch Learning)**

Attention	$P(a\|w)$	$P(w\|a)$	[8]
VC1 (touch hand)	0.4	**0.8**	0.4
VC2 (touch hand + next to)	0.6	0.5	**0.8**
VC3 (eye-gaze)	0.2	0.2	0.2
VC4 (all combined)	0.2	0.4	**0.8**

[8] C. Yu and D. H. Ballard. A unified model of early word learning: Integrating statistical and social cues. Neurocomputing, 70(13):2149-2165, 2007.

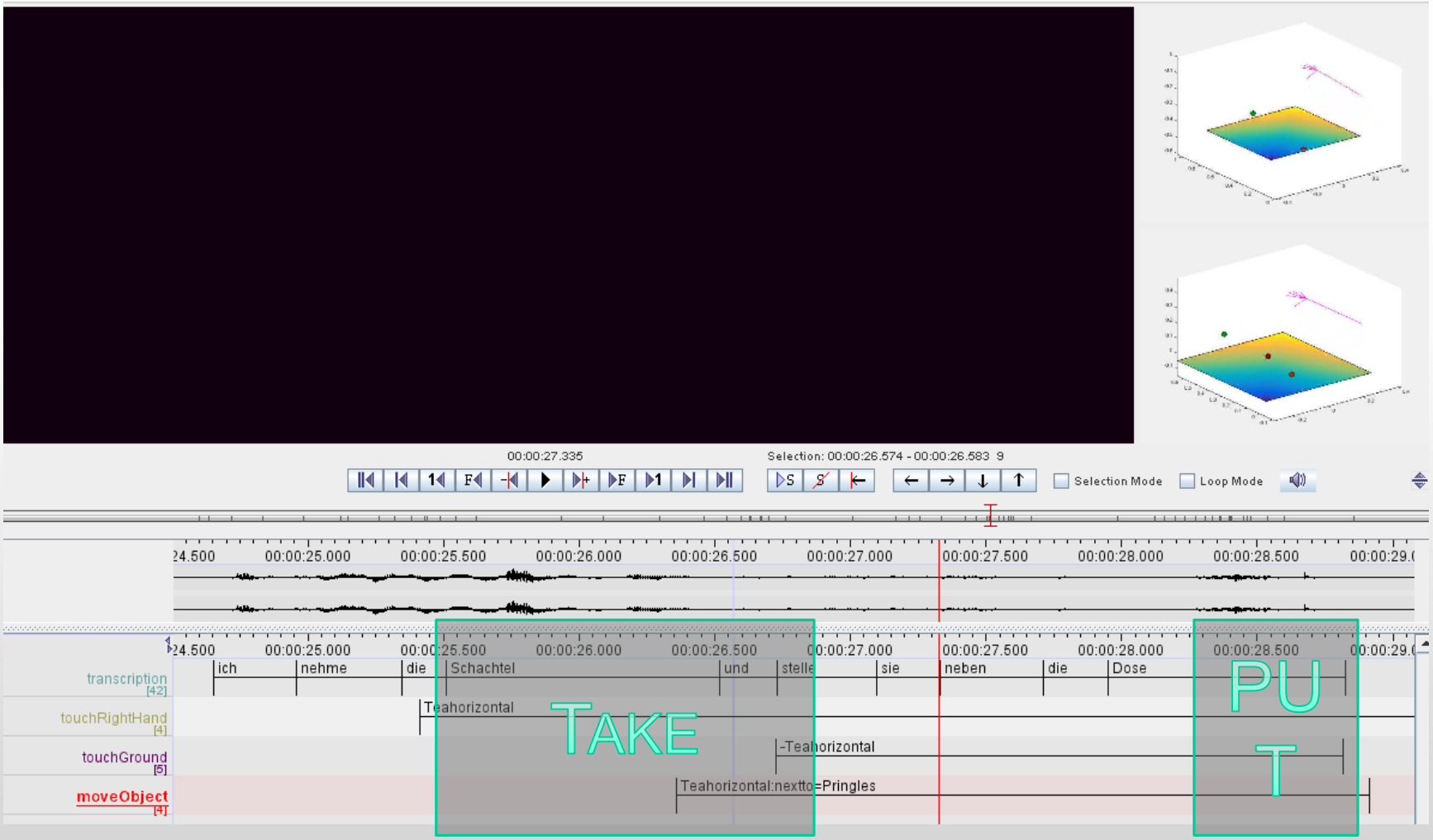
Crossmodal word learning

- Early stage word learning: associations between
  - words (perhapts morphologically simplyfied – lemmata??)
  - references to:
    - objects in the environment
    - object types
    - object classes (hypernyms)
- Action vs. object references:
  - children for some time have a strong bias towards either learning preferredly nouns first, or verbs.
  - which is preferred is culture/language specific (Gogate & Hollich 2016)

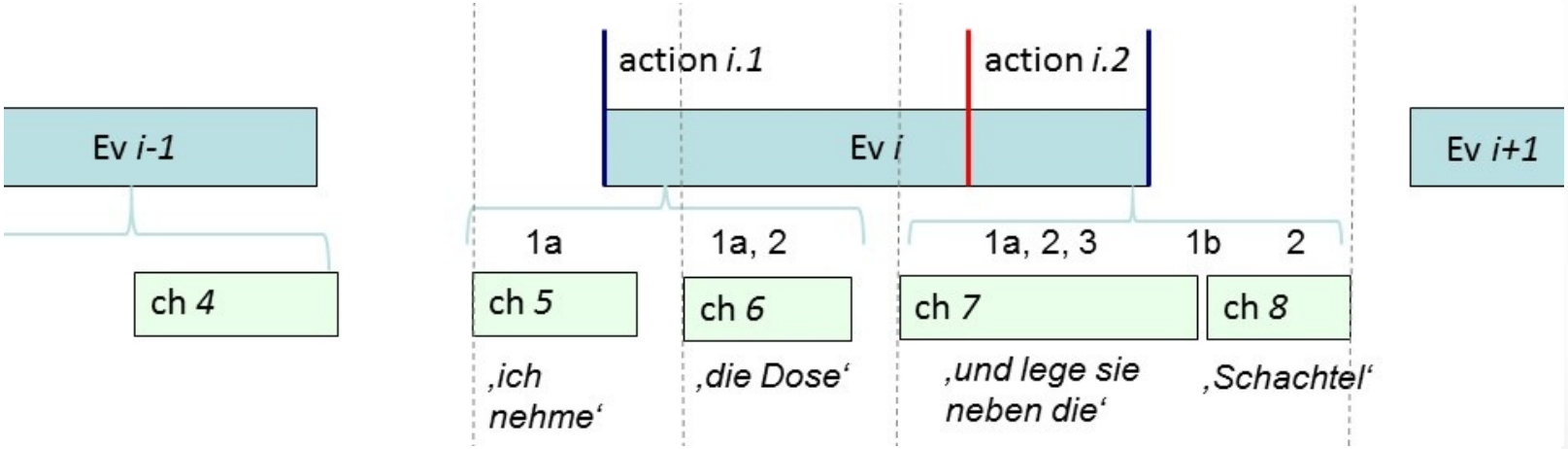
A pipeline for incremental word learning

- Object references are given by the objects themselves. If an object is in the focus of attention has to be determined by specific cues
- Actions involving particular objects put the focus on these objects
- What we need is therefore:
  - segmentation and identification of actions
  - identification of objects involved
  - alignment between actions and speech describing the action
  - an algorithm that extracts specific word-reference pairs out of the multiple streams of multimodal input data

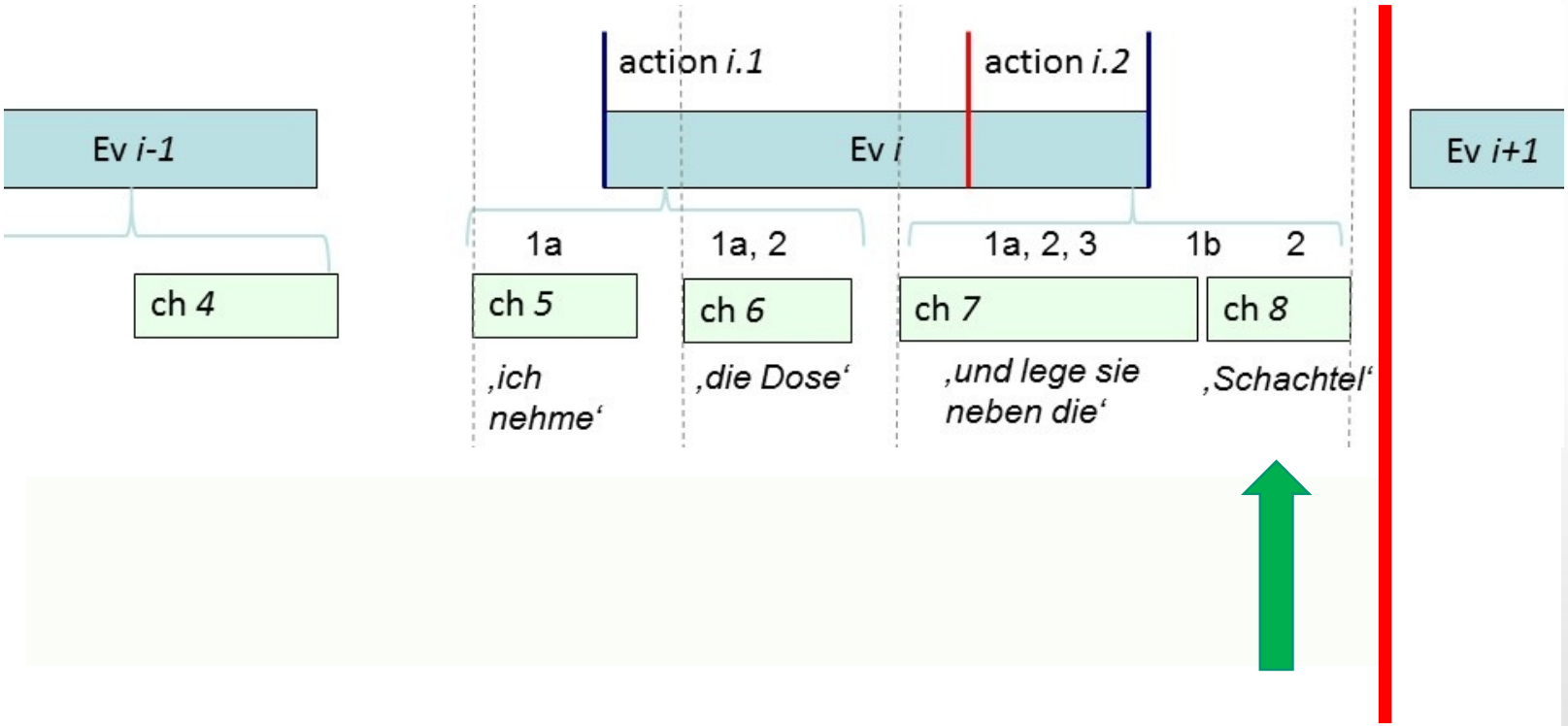
Action Segmentation



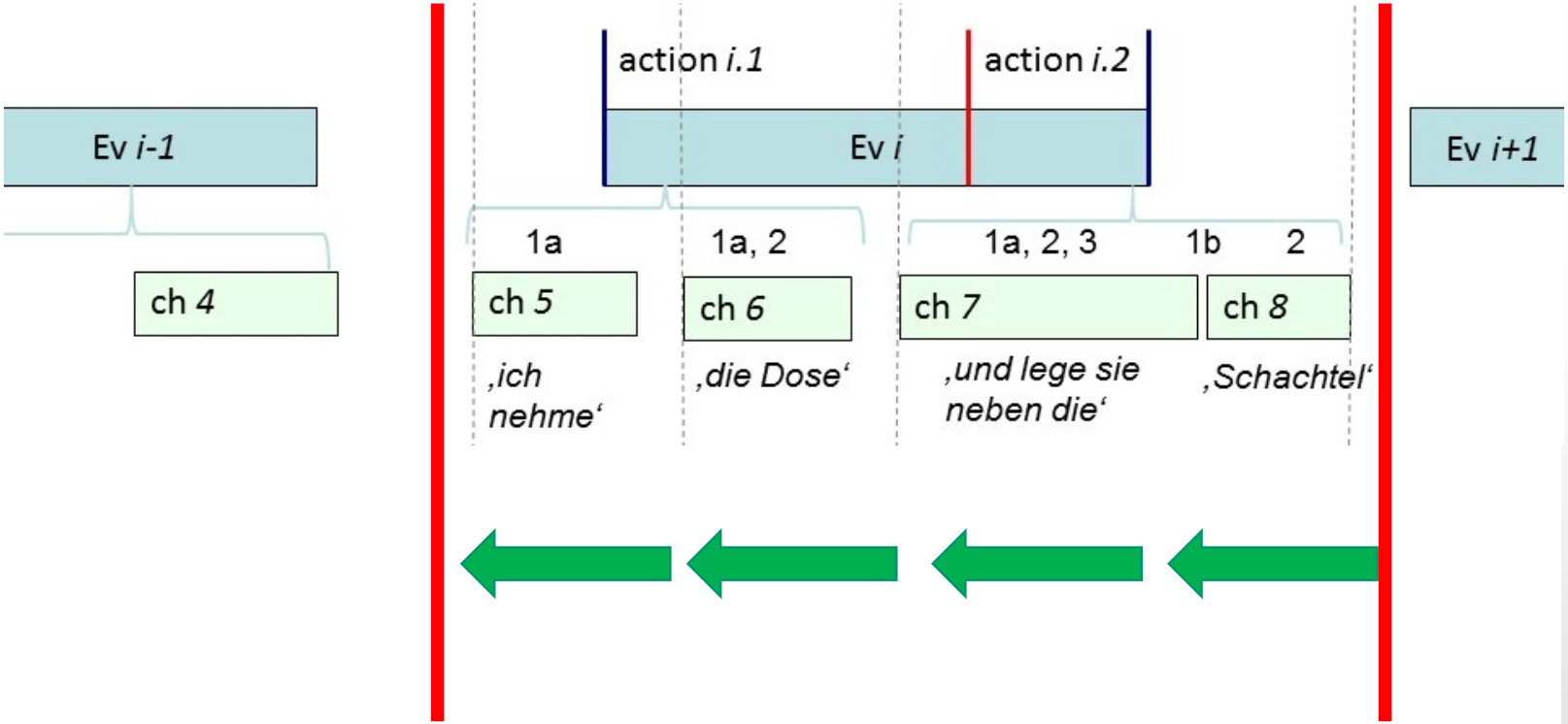
Alignment of Actions and Speech



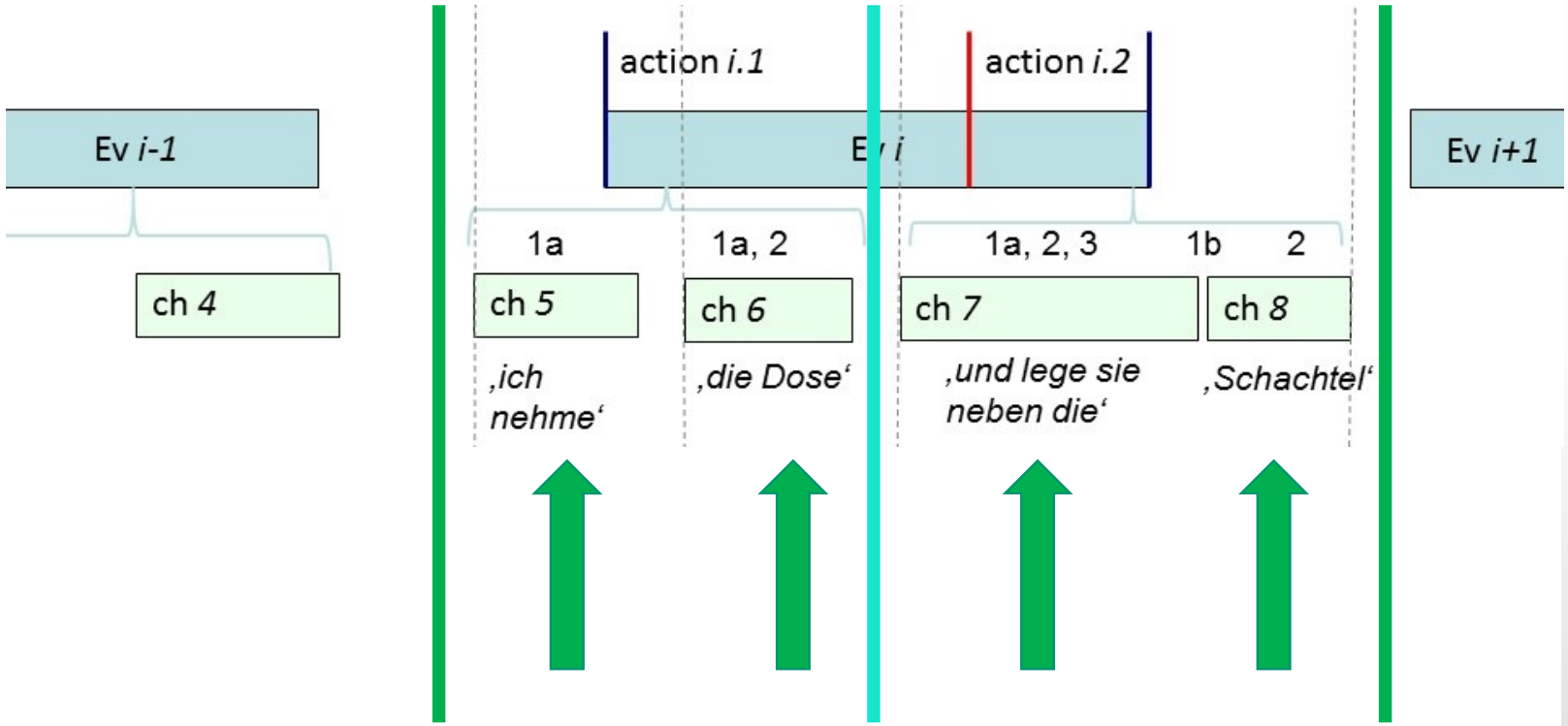
Alignment of Actions and Speech



Alignment of Actions and Speech



Alignment of Actions and Speech

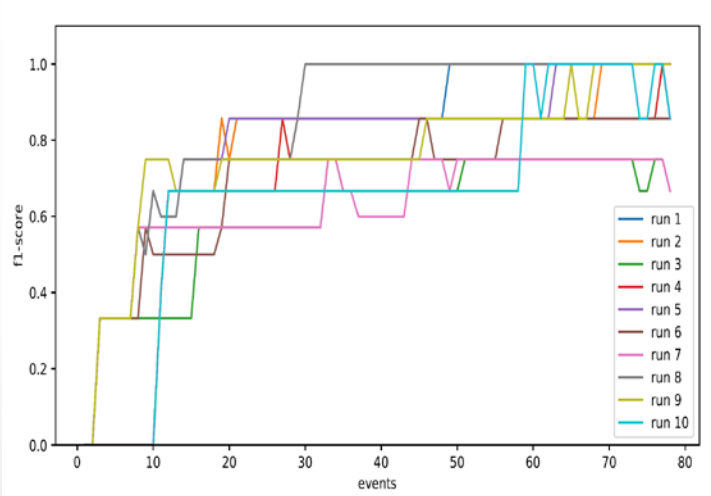


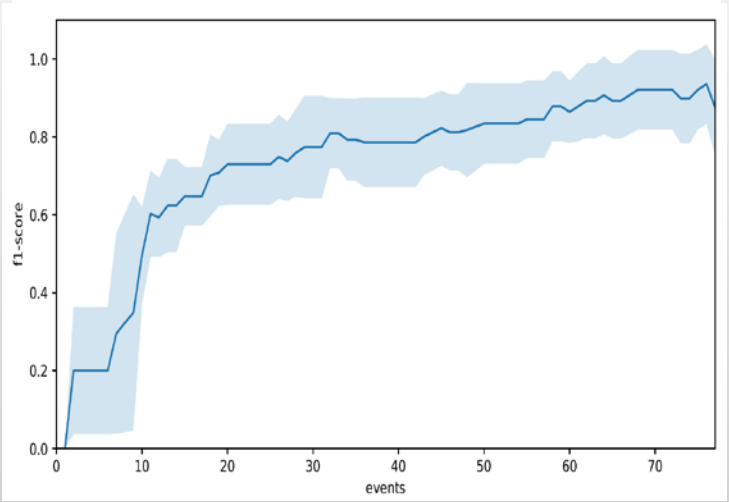
Incremental Statistical Lexicon Learning

- A simple algorithm that incrementally learns word-reference mappings.
- References can be to objects and actions.
- At each step, for each potential mappings, statistical information is gathered: pmi, npmi, p(w|r), p(r|w).
- The most reliable value is npmi (normalized pointwise mutual information).
- Mappings are ranked to each other and compete with each other.
- If a mapping is ranked high and meets certain positive thresholds, it is included in the lexicon.
- If a mapping in the lexicon is ranked lower than a competing one and meets certain negative thresholds, it is removed from the lexicon.

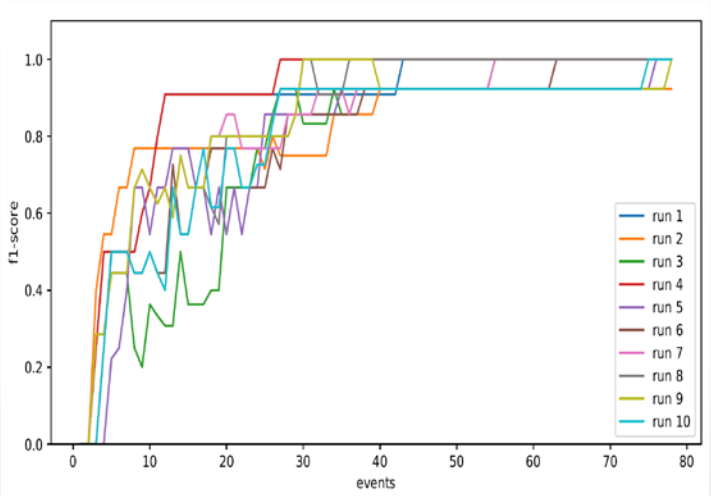
Results

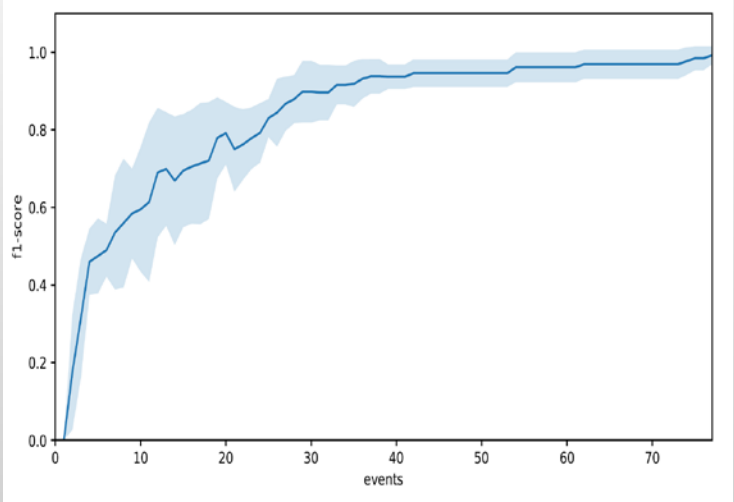
Word/Action Mapping (full forms)





Word/Action+Object Mapping (full forms)





Multimodal Lexicon Learning: Architecture

- Input channels:

Visual input:

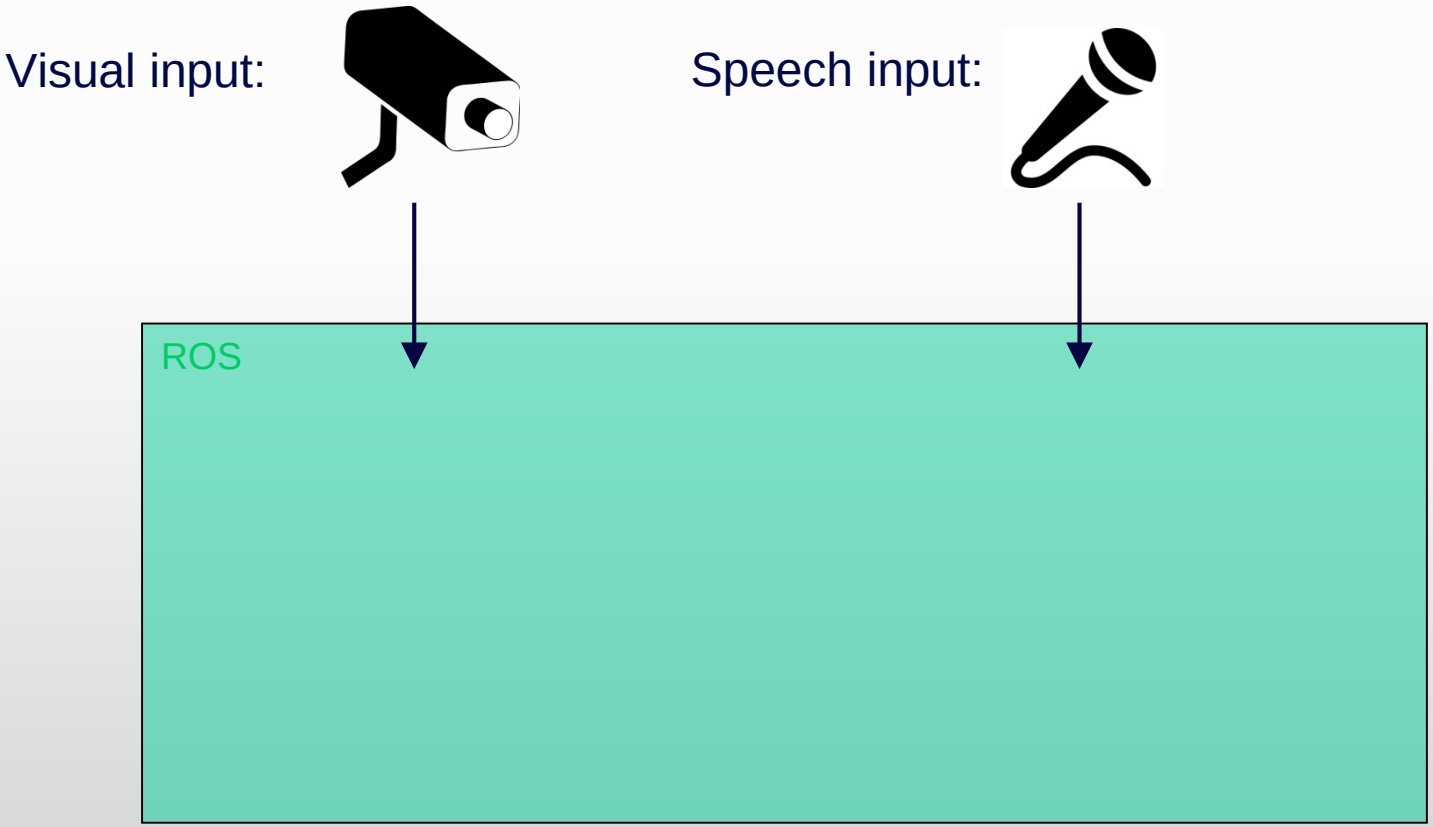
Pepper‘s camera

Speech input:

Google ASR

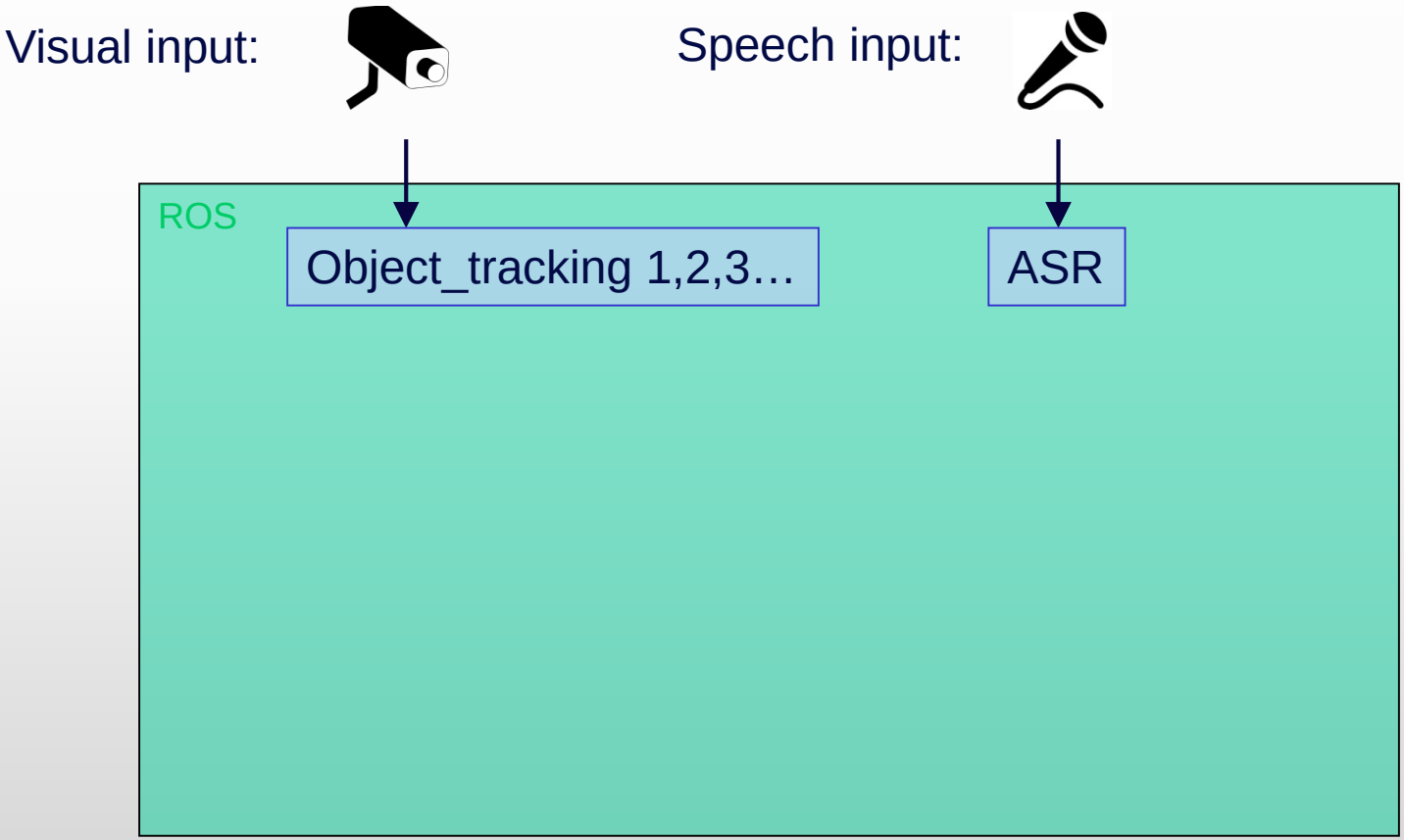
Multimodal Lexicon Learning: Architecture

- ROS (robot operating system): almost all functionality



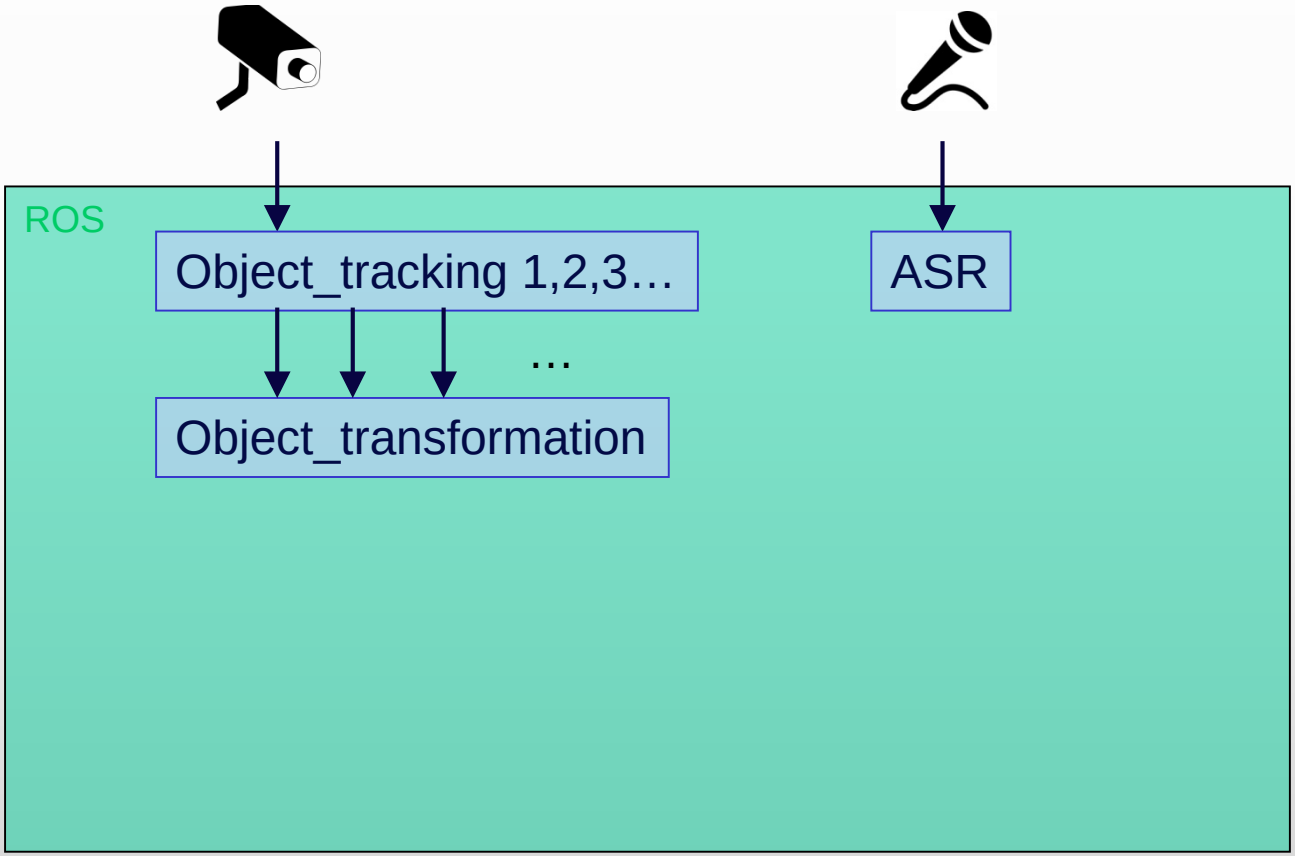
Multimodal Lexicon Learning: Architecture

- Object tracking / speech recognition:



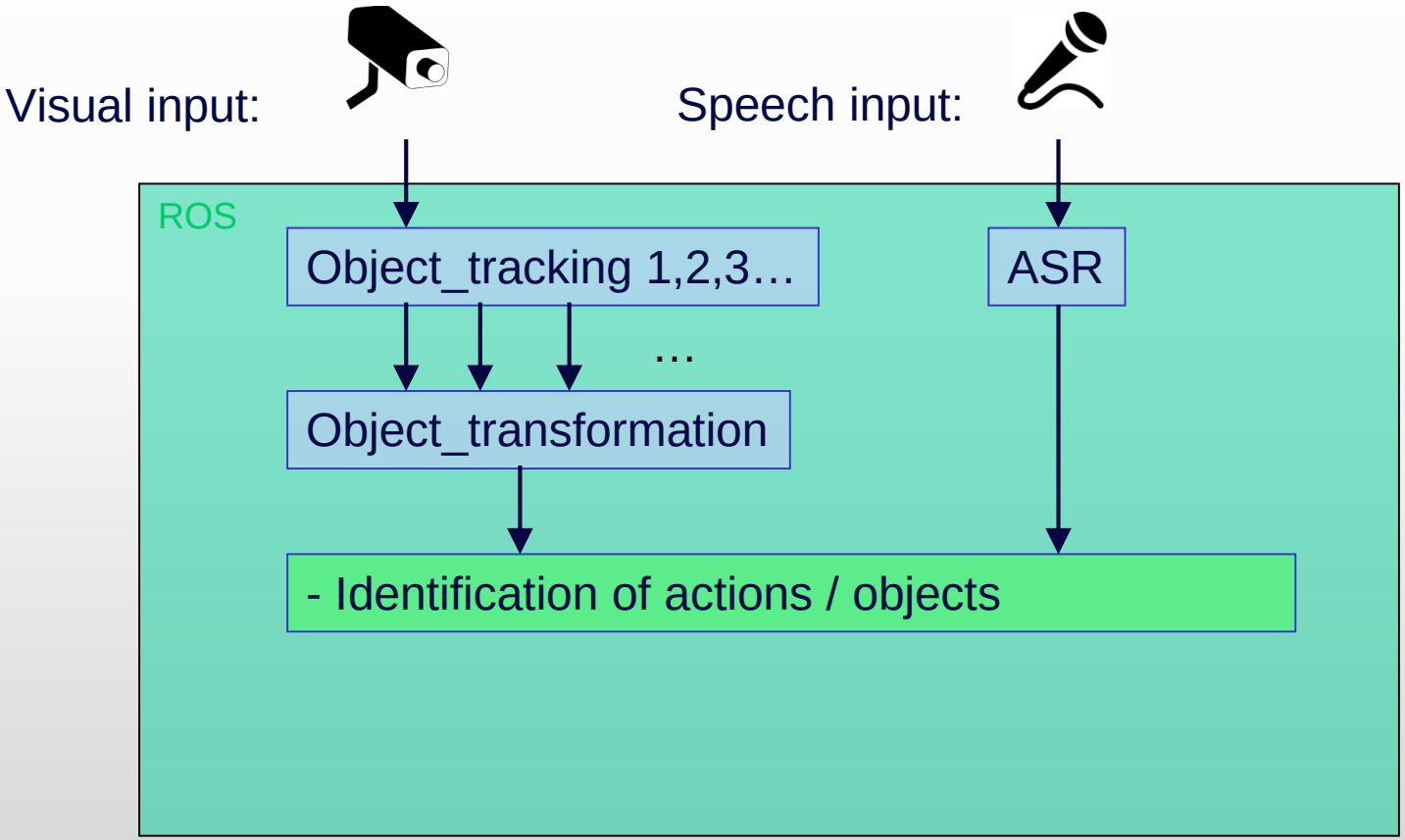
Multimodal Lexicon Learning: Architecture

- Transformation of coordinates to „table“:



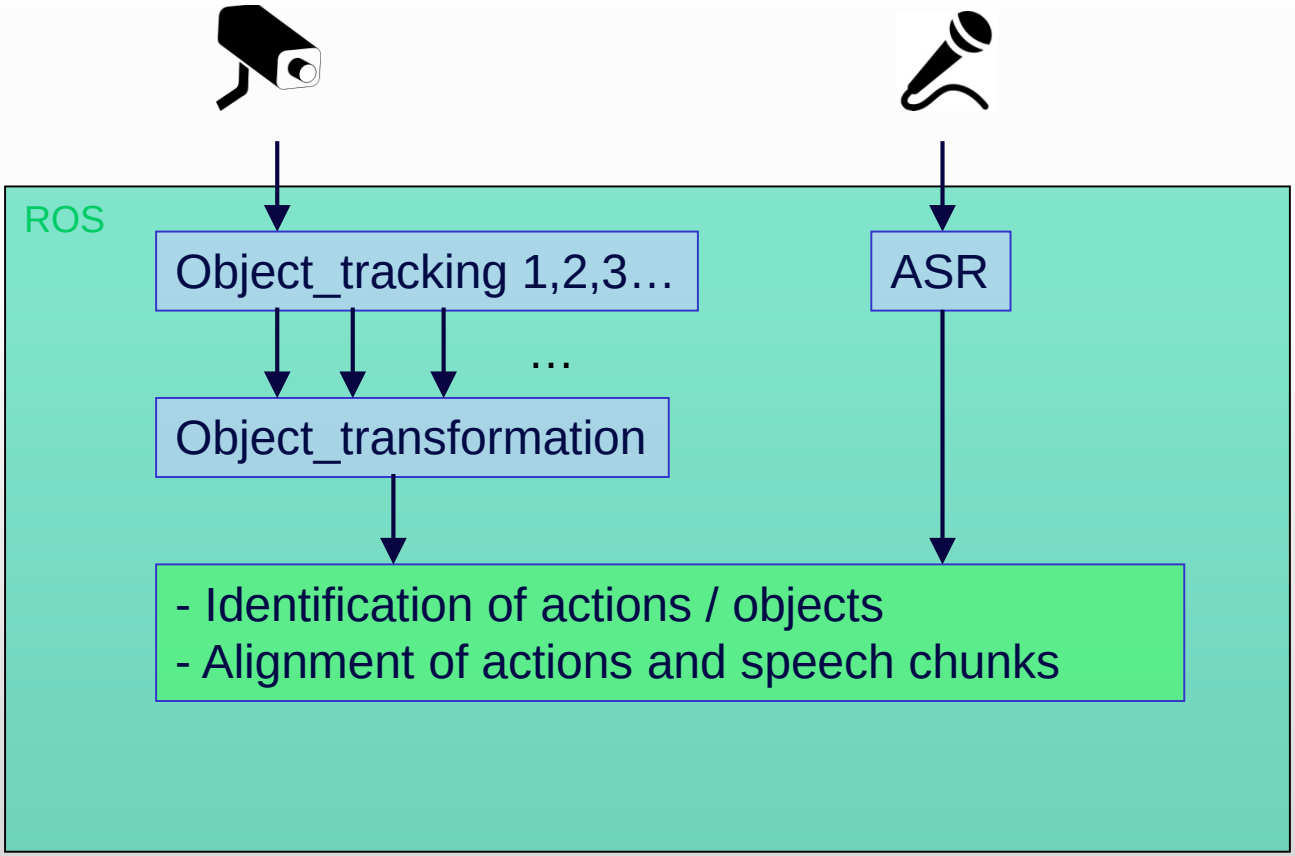
Multimodal Lexicon Learning: Architecture

- Multimodal integration:



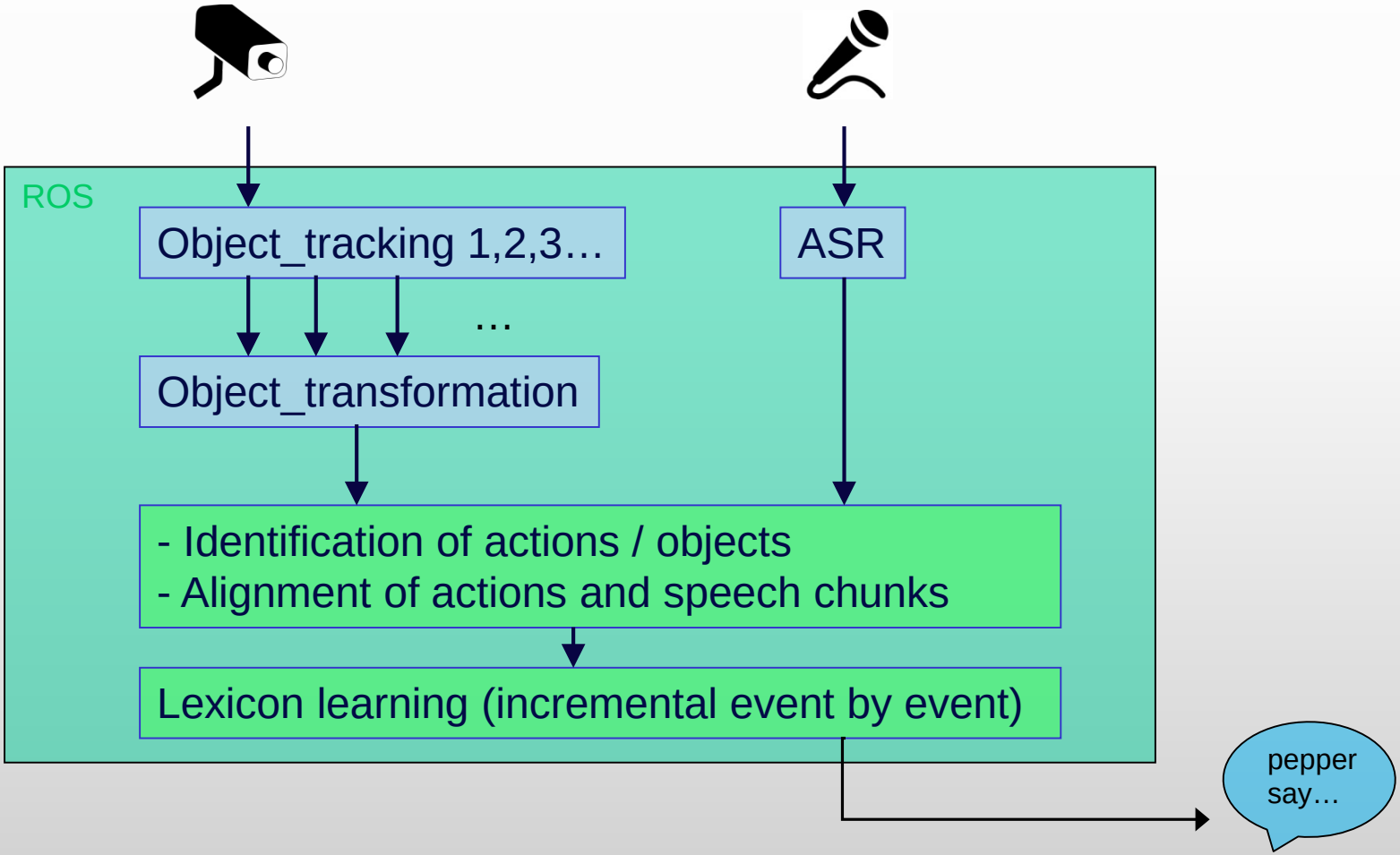
Multimodal Lexicon Learning: Architecture

- Alignment of speech and visual input:



Multimodal Lexicon Learning: Architecture

- Lexicon learning: Incremental Information Theoretic Model



Thank you for your attention!

CHIST-ERA HLU
Project ATLANTIS
http://atlantiscom.wordpress.com

WWTF
Viennese Science and Technology Fund
Project RALLI
http://ralli.ofai.at

Data Sets

- OFAI Multimodal Task Description Corpus (MMTD)
  http://ofai.at/research/interact/MMTD.html
- Action Word Corpus (AVC)
  http://ofai.at/research/interact/avc.html

!!! Upcoming Workshop !!!
ICMI 2018, Boulder, Colorado
October 16th, 2018
Workshop on Cognitive Architectures for Situated Multimodal Human Robot Language Interaction
Paper deadline: 29. June 2018
http://ralli.ofai.at/workshop.html